Title:

An investigation of the inter-rater reliability of the Valpar Joule Functional Capacity Evaluation in healthy adults.

ABSTRACT

Background

A functional capacity evaluation (FCE) can provide a comprehensive, objective measure of a worker's ability to meet work demands to support return to work decision making. Research evidence of a FCE's reliability and validity, involving more than one study, and covering all test components with a diverse range of populations, is essential to ensure confidence in any FCE system.

Objective

This study aimed to establish the inter-rater reliability of the Valpar Joule FCE functional capacity evaluation (FCE) for which there is currently limited published literature regarding its reliability.

Methods

Twelve healthy subjects were digitally recorded completing the initial protocol of the Valpar Joule. Assessments were rated separately by 3 raters and the results then compared.

Results

Using Intraclass Correlation Coefficients (ICC), with percentages of agreement and t-tests to determine bias, inter-rater reliability was high for determining last safe weight lifted for forceful tasks with ICC'>0.90. Agreement ranged from 97.2%-100% for determining reasons for terminating tests; 97.2%-98.6% for identifying maximum safe capacity, but was only between 8.3% – 50% for full agreement for identification of last weight safely lifted in forceful tasks. Differences were identified between raters with different training and experience for identifying poor body mechanics in lifting.

Conclusion

Results demonstrated high inter-rater reliability for the Valpar Joule functional capacity evaluation in healthy adults. Further development of criteria identifying poor body mechanics and training in its use is recommended to increase evaluator objectivity.

Key words: return to work; physical capacity; body mechanics.

1. INTRODUCTION

Sickness absence from work is prevalent in the UK and costs around £100 billion annually in "sickness benefit" payments [1,2]. Current Government policy has a strong focus on reducing sickness absence and improving services to promote work health and wellbeing [3, 4]. While employers may wish absent employees to return to work (RTW) quickly, they may require detailed healthcare guidance on an employee's abilities, performance limitations or necessary work adjustments in order to support safe RTW [1,5].

Allied Health Professional's (AHPs) and General Practitioners (GPs) can help to provide employers with this information (Allied Health Professions Federation [2, 5-7]. However, objective information is required to support any recommendations of when and how absent workers can return safely [8]. A common component of work rehabilitation programmes which strives to provide objective data is the Functional Capacity Evaluation (FCE) [9]. FCEs are test batteries which provide a systematic, comprehensive and performance based measurement to determine an individual's physical abilities to carry out work related tasks [10].

While it is not suggested that an FCE be used in isolation, studies have identified they can make a significant contribution to a comprehensive RTW process, by comparing the injured worker's physical capacity with their work demands to determine ability to RTW safely [11-13]. However, evidence of FCEs reliability is vital to demonstrate that changes over time in a worker's performance are due to variation in their abilities and not as a result of FCE measurement error and ensure there is confidence in the assessment results [14,15]. Methodological limitations in studies supporting FCE reliability have previously been highlighted [14, 16]. Therefore, while it is acknowledged that there are a number of other FCE's available, this study aimed to further evaluate the inter-rater reliability of one FCE, the Valpar Joule Functional Capacity Evaluation [17, 18] and to identify any factors potentially effecting inter-rater reliability.

2. METHOD

2.1 Research Design

A cross-sectional study was conducted to investigate the inter-rater reliability of three raters conducting the Valpar Joule Functional Capacity Evaluation.

2.2 Participants

Gouttebarge et al [16] rate studies involving more than 2 raters with 10 or more subjects highest in their systematic review of rater reliability of FCEs. Therefore a convenience sample of 12 healthy male and females of working age (18-65 years) was recruited from University staff and associates.

Subjects were screened using PARQ [19] and were excluded if they had current or chronic injury or illness; elevated resting Blood Pressure >180/90; or elevated Heart Rate >(220-clients age)x 0.85 according to Valpar Joule procedural guidance [17].

2.3 Raters

Three raters -1 primary and 2 silent raters, carried out the FCE assessments. The primary rater (rater 1) and a silent rater (rater 3) were both experienced occupational therapists who had completed an approved Valpar Joule training programme and had similar FCE experience (5 years). Due to the lack of another local Valpar Joule FCE rater, the other silent rater (rater 2) was an experienced physiotherapist with expertise in biomechanics and musculoskeletal rehabilitation but with no FCE rating experience.

2.4 Procedure

Ethical approval for this study was obtained from the University School Research Review Group.

Subjects were recruited via email that included an information pack outlining the study and provided written informed consent prior to their participation. Subjects attended a one-off session and had initial demographic details recorded (age, sex, height, weight, and work status). Blood pressure (BP) and

heart rate (HR) were recorded prior to and after assessment, with HR monitored during assessment, to ensure these did not exceed safe limits [17].

Each subject completed the V.J. FCE initial protocol [17], directed and rated by the primary rater, requiring the subject to complete a series of:

- 8 forceful tasks (waist to waist lift; waist to floor lift; waist to above shoulder lift; bilateral carry; unilateral dominant hand carry; unilateral non-dominant carry; push; and pull tasks). Due to push and pull tasks being tested using a force gauge which produces an absolute number and therefore 100% agreement these results were not discussed in this study and subsequently only 6 forceful tasks will be reported on.
- 6 positional tasks (sitting; standing; kneeling; crouch; sustained mid-level reach; sustained elevated reach).
- 8 repetitive tasks (walking; crawling; stair climb; ladder climb; balance; repetitive foot (right and left); fine motor co-ordination).

The primary rater observed each task completion, determining the last safe weight the subject lifted by identifying unsafe body mechanics or physiological signs; the reason each task was terminated; and determining the subject's maximum safe capacity. Following studies by Legge and Burgess–Limerick [20] and Reneman et al [21], the assessment was digitally recorded using video cameras from more than one angle with sound recording to maximise the "silent" raters view of a subjects performance and awareness of reported relevant factors such as participant's cardiopulmonary function recordings, perceived rate of exertion [22] or any subjective comments. Recordings were transferred to external hard drives and the two "silent" raters independently viewed and rated each subject's assessment blinded to each other's and the primary rater's ratings reducing the likelihood of rater bias [23].

All test data was coded, stored and analysed in accordance with Robert Gordon University policy and the Data Protection Act [24] to ensure confidentiality and protect the anonymity of volunteers.

2.5 Data Analysis

For each subject, the FCE results for each rater were statistically analysed to investigate inter-rater reliability and issues impacting on reliability. Comparison between all raters and pairs of raters was made to identify consistency and agreement or differences between raters. Analysis was performed using statistical analysis software Statistical Packages for Social Sciences (SPSS) for Windows [25].

The level of inter-rater reliability for determining last safe lift in the forceful tasks of the protocol producing ordinal data was calculated using a two way mixed model intraclass correlation coefficient (ICC 2,1) to determine absolute agreement [26,27]. To allow for comparison with other FCE studies, the levels of reliability scale provided by Gouttebarge et al [16] in their review of FCE's, was adopted and levels of reliability were determined as: High for ICC >.90; Moderate 0.75<ICC>0.90; Low<0.75.

It has been acknowledged that no single test provides a complete measure of reliability [28, 29]. Therefore, in addition to the ICC (2,1A), a 95% confidence interval (CI) was calculated for each ICC mean. Percentages of agreement were also used to determine agreement between all raters and pairs of raters [14]. Ranking was used to identify the number of ties or positive or negative ratings between pairs of raters [27] and t-tests completed to determine rater bias [26].

Percentages of agreement were also used for comparing inter-rater reliability in raters scoring for tasks which produced nominal data for:

- Reasons for stopping all tasks in the Valpar Joule protocol (rated 1-3 as: 1. Task was fully completed; 2. Subject determined to be using unsafe body mechanics; or 3. Subject exceeded maximum safe exertion levels according to VAI guidelines [17] and as determined by physiological measures.
- Identifying the maximum safe capacity of subjects in each of the forceful, positional and repetitive tasks in the Valpar Joule protocol (17) (rated 1-3 as: 1. No identified limitations; 2. Occasional; or 3. Rare).

3. RESULTS

3.1 Subject Demographics

A convenience sample of 12 healthy subjects (8 women, 4 men) was recruited. All subjects were employed either in teaching (n=10) or manual work (n=2). Subject ages ranged from 18-59 years with a mean of 40.58 years and standard deviation (SD) 12.5 years. Height ranged from 150cm -183cm with mean of 168.33cm and SD of 10.34cm. Weight ranged from 56 kilograms (kg) to 98kg with a mean of 76.25kg and SD of 13.90 kg. Each subject completed all aspects of the protocol in the determined order, apart from one subject (subject 1) who was unable to complete the push task due to equipment failure.

3.2. Determination of last safe weight in Forceful Tasks

3.2.1 Intraclass correlation coefficients – all raters

The level of inter-rater agreement for all forceful tasks was high with all ICC>0.9 and narrow CIs with the largest interval ranging from 0.738- 0.987 for unilateral non-dominant carry and the narrowest interval of 0.939-0.997 for waist to floor carry (table 1).

3.2.2 Percentages of agreement – all raters

While high ICC > 0.9 are reported, for the total 72 possible scored lifts (6 lifts per 12 subjects), it is interesting when looking at the actual % of agreement to note that all 3 raters only fully agreed on a subject's last safe weight lifted in 15/72 lifts (20.83%). Highest agreement was achieved between all 3 raters in bilateral carry for 6/12 subjects (50%). Full agreement was lowest on waist to floor, waist to above shoulder lifts and unilateral dominant hand carry when achieved for 1/12 subjects (8.3%). In contrast, highest full disagreement was recorded for 4/12 subjects (33.3%) in waist to floor lifts (Table 1).

Due to the difference in each incremental weight being lifted, differences in the ratings of last safe weight lifted recorded by the raters were identified being up to 4 incremental weights in some lifts

which equated to between 3 pounds of force for waist to floor lift and 16 pounds of force for bilateral, unilateral dominant and non-dominant hand carry.

Insert table 1 here

3.2.3 Intraclass Correlation Coefficients - Paired Raters

In determining last safe lift, inter-rater reliability was determined as high for each pair of raters with all ICC >0.90 and the narrowest confidence interval ranging from -0.081 - 0.414 for waist to waist score for pair 2 (raters 1 and 3) (Table 2).

3.2.4 Percentages of agreement - Paired Raters

Table 2 reports that the highest percentage of agreement for determining last safe weight lifted was identified between raters 1 and 3 who agreed on 45/72 (62.5%) lifts. Agreement was identified as lower between other pairings, with paired raters 1 and 2 agreeing in 24/72 (33.3%) last safe weight scores and paired raters 2 and 3 agreeing on 22/72 (30.5%) scores. Paired raters 1 and 3 achieved the highest agreement for determining last safe weight lifted for any task in waist to waist lift when they agreed in 10/12 (83%) of subjects' ratings.

3.2.5 Paired Differences

Table 2 also identifies that there was a significant difference identified for rater 2 scores compared to raters 1 and 3 in determining last safe weight lifted in 4/6 (66.7%) of the forceful tasks – in waist to waist, waist to floor, waist to above shoulder lifts and unilateral dominant hand carry. No significant difference in scores was noted between raters 1 and 3.

A difference of up to 4 incremental last safe weights lifted by a subject was recorded between raters, equating to a difference of up to 16lbs of force being reported as the safe lifting ability for a subject in some lifts.

Insert table 2 here

3.3 Reasons for terminating tasks and maximum safe capacity in forceful, positional and repetitive tasks

A high percentage rate of agreement was identified between all raters for both determining reasons for terminating tests (table 3) and for identifying maximum safe capacity (table 4) for each of the 12 subjects completing all 20 tasks in the protocol (240 tasks in total as push and pull not included): 100% (72/72 tasks) agreement for forceful tasks; 97.2% (70/72 tasks) agreement for positional tasks; and 98.6% (95/96 tasks) for repetitive tasks. The maximum safe capacity for forceful tasks is not presented in the results as this was calculated from the score each rater gave for last safe weight lifted and results would be the same as those presented for last safe weight.

It was noted that where there was difference in ratings for terminating tests, this was where rater 3 recorded 3 subjects as having completed a task while the other raters accurately reported that 2 subjects had their test stopped due to unsafe body mechanics and 1 subject had asked for the test to be stopped (table3). This subsequently had an impact on the raters' determination of these subjects 'maximum safe capacity as rater 3 rated each of those subjects as having no limitations, although tests were stopped due to the reasons reported (table 4). This might therefore be considered as recording errors and should be taken into consideration when viewing these inter-rater reliability results.

Insert Table 3 then table 4 here

4. DISCUSSION

4.1 Inter-rater reliability

The aim of this study was to evaluate the inter-rater reliability of the Valpar Joule (V.J.) FCE and to consider any factors which effect reliability or its use as a clinically effective FCE tool. To allow comparison with other FCE inter-rater reliability studies, methodology as determined by Gouttebarge et al [16] for determining inter-rater reliability was adopted. The findings of this study identified that inter-rater reliability for determining the last safe weight lifted for each forceful task subtest of this FCE

protocol [17] was high as evaluated by ICC >0.90 and with narrow confidence intervals, ranging from 0.738- 0.987 for unilateral non-dominant carry to 0.939-0.997 for waist to floor carry (table 2). Reasons for terminating tests and identifying maximum safe capacity were also identified as having high inter-rater reliability, as determined by percentages (%) of agreement, ranging from to 97.2%-100% for agreement for reasons for terminating tests and from 97.2%-98.6% for identifying maximum safe capacity.

However, when the actual raters' scores for determining the last safe weight lifted by each subject were analysed using percentages of agreement, it was apparent that there was some significant difference in agreement between raters. It was identified that all raters only fully agreed with each other's ratings on 20.83% of occasions (15/72) and also fully disagreed on a similar amount of occasions, 15.3% (11/72). Raters achieved full agreement most frequently for bilateral lift ratings (50% or 6/12) but lowest for waist to floor, waist to above shoulder and unilateral dominant hand carry when full agreement was only achieved in each task on one occasion (8.3% or 1/12).

Some significance difference was noted between pairs of raters. It was identified that rater 2 only agreed with the raters 1 and 3 in 30.5% - 33.3% respectively of total scores. Rater 2 also recorded significantly higher weights which subjects could safely lift in most lifts and a significant difference was identified in their scores compared to raters 1 and 3 in determining last safe weight lifted in 4/6 (66.7%) of the forceful tasks – in waist to waist, waist to floor, waist to above shoulder lifts and unilateral dominant hand carry.

This was in contrast with paired raters 1 and 3 who were in full agreement in over 62.5% of their total scores but with a range between 33.3% and 83.3%. This is almost twice the level of agreement than was identified between their pairings with Rater 2 and there was no significant difference identified between their ratings of last safe weight lifted in any lifting task.

4.2 Evaluator Training

The impact of evaluator discipline or training on FCE ratings has been recognised as being relatively unknown [30]. Therefore while this study's primary aim was not to investigate the impact of training or discipline on reliability of the VJ FCE, it is of interest to note that raters 1 and 3, were both

occupational therapists with similar background; FCE training; experience in safe return to work decision making and that they showed no significant difference in their ratings when compared with rater 2, an experienced physiotherapist with extensive musculoskeletal and biomechanics experience but no V.J. FCE evaluator training. Of interest, raters 1 and 3 also significantly scored subjects as having lower lifting capacity than rater 2. Subsequently, while the importance of an evaluator's skill level is recognised, these findings would appear to support the view that an evaluator discipline, knowledge, training and experience can impact on an evaluator's judgements, confidence, objectivity of test scoring and subsequently on inter-rater reliability of a FCE [30-33]. Psychosocial, environmental and cultural factors may also influence evaluator's judgements [34, 35]. Therefore, it is suggested that these factors and the evaluator's clinical reasoning when conducting the FCE, would all benefit from further investigation to increase the accuracy of assessment recommendations and support sustained return to work in order to increase the inter-rater reliability of the Valpar Joule.

4.3 Criteria for determining safe body mechanics

While low percentages of full agreement between raters are reported in this study, it should be noted that when the raters were not in absolute agreement, they only differed by a small number of incremental weights when determining the last safe lift completed. Most frequently raters only differed by 1 incremental weight but on occasion there were differences of between 2-4 incremental weights. Unfortunately, due to the incremental weight differences in lifting and carry tasks, depending on which work level [17] a subject was lifting at, in some cases the difference in the rater's decision regarding a subject's last safe weight lifted was 16 pounds of force. It should then be considered that it may be likely that this difference in weight could have significant impact on whether an individual was determined to meet the demands of their job and able to return to work safely. Therefore it is essential for both employees and employers that improvements are made in evaluator's observations and criteria for determining last safe weight lifted in this FCE in order to most accurately reflect a worker's lifting abilities, minimise any difference in ratings and to facilitate safe work return.

It could be suggested that this result reflects the views of Tuckwell, Straker and Barrett [36], who highlighted that lifting is the most subjective part of the FCE and requires determination of quality of posture and movement. It is acknowledged that one of the raters in this study, while an experienced

physiotherapist, was not trained in the use of this FCE and therefore could have been expected to score significantly differently from the two trained raters. However, differences were also reported in percentages of agreement in all lifts between the two trained raters and their subsequent determinations of a workers safe lifting ability. While evaluator's clinical reasoning in FCE decision making has been identified as requiring further investigation [31], it has been identified that defining safe maximal lift [37] and developing sound FCE rating criteria could reduce subjectivity in testing and that training on how to interpret criteria, and consistent application of a rating scale can enhance objectivity [38]. King, Tuckwell and Barrett [33] concur, noting that objectivity can be promoted when procedures, variables for observation and scoring are all operationally defined. When reviewing which lifts achieved the most agreement or disagreement between raters, no particular pattern was established. Rater differences were not particular to any subject being tested and no specific explanation for why there was greater agreement in some tasks was identified. Therefore, it is suggested that the rating criteria for determining poor body mechanics for the VJ FCE should be developed to enhance inter-rater agreement on last safe weight lifted in forceful tasks to minimise any discrepancies in the FCE results and subsequent return to work recommendations. This will help ensure the confidence of raters in their assessment results and subsequent recommendations [13].

4.4 Limitations

This small study used a convenience sample of healthy young adults and results cannot be generalised to the wider population or to individuals with specific health conditions [39]. Additionally the effects of subject gender, weight or height were not taken into consideration in this study. Further research is now necessary to establish inter-rater reliability of the V.J. FCE with injured workers or individuals whose conditions can change [31, 40,41]. Future research is also required to establish other forms of reliability and validity of the V.J. FCE to ensure there is confidence in the assessment results [14].

It is acknowledged that due to lack of availability, inter-rater agreement for the V.J. FCE could not be determined between 3 V.J. trained evaluators. However, in using an experienced physiotherapist alongside two V.J. trained evaluators, the high inter-rater reliability for aspects of the FCE was still determined and the value of the training for evaluators apparent. It would have also been of interest to

involve a third Occupational Therapist untrained in V.J. evaluation to provide a comparison of reliability, based on the V.J. training.

5. CONCLUSION

This study investigated the inter-rater reliability of the Valpar Joule FCE and identified high inter-rater reliability for lifting and carrying tasks determined with intraclass correlation coefficients of >0.90 and a high percentage of agreement between all raters of > 90% for reasons for terminating tests and identification of a subject's maximum safe working capacity. However, the findings also revealed a significant difference in scoring between pairs of raters for identifying last safe weight lifted in forceful tasks. The study highlighted apparent differences in rater's views on criteria for determining poor body mechanics.

It appears that different training and experience may impact on objectivity of test scoring and subsequently on inter-rater reliability [30-33]. While it is acknowledged that the Valpar Joule provides training and suggested criteria, given the findings of this study and consequences of a FCE results, it is concluded that the objectivity of observations for lifting and carry tasks and the inter-rater reliability of the V.J. FCE could be further enhanced. Further consideration of factors which may improve objectivity is required and the development of more specific, clearly defined criteria for determining the presence of physical signs of poor body mechanics and additional rater training to assist in their detection is also recommended to improve rater skills, objectivity, minimise discrepancies in ratings and increase confidence in results for the V.J. FCE.

It was also recognised that the evaluators' clinical reasoning when conducting FCEs and the effect of an evaluator's experience, training, and incorporation of information other than just biomechanical factors would benefit from further investigation [31-33,38]. How this subsequently impacts on assessment recommendations and successful, sustained return to work is also suggested as an area for future research to increase the accuracy of assessment recommendations.

Acknowledgements:

[Edited for the review process]

Conflicts of interest:

There was no conflict of interest for all authors in this study.

6. REFERENCES

1. SCOTTISH GOVERNMENT. Health works: a review of the Scottish governments Healthy Working Lives strategy. [online]. Edinburgh: The Scottish Government; 2009. Available from: http://www.scotland.gov.uk/Publications/2009/12/11095000/4 [accessed 11th February 2012].

2. DEPARTMENT OF WORK AND PENSIONS. Improving Health and work: changing lives. [online]. London: The Stationary Office; 2008. Available from: http://www.workingforhealth.gov.uk/documents/improving-health-and-work-changing-lives.pdf [accessed 25th November 2008].

3. DEPARTMENT OF WORK AND PENSIONS. Policy Publications. [online]. London: The Stationary Office; 2012. Available from: http://www.dwp.gov.uk/publications/policy-publications/ [accessed 11th September 2012].

4. SCOTTISH GOVERNMENT. Health and work. [online]. Edinburgh: The Scottish Government; 2012. Available from: http://www.scotland.gov.uk/Topics/Health/Healthy-Living/Health-Work [accessed 11th February 2012].

5. BLACK, C., & FROST, D. Health at work: an independent review of sickness absence. [online]. London: The Stationary Office; 2011. Available from: http://www.dwp.gov.uk/docs/health-at-work.pdf [accessed 11th September 2012].

6. ALLIED HEALTH PROFESSIONS FEDERATION. Allied Health Professions Advisory Fitness for Work Report, [online]. London: Allied Health Professions Federation; 2013. Available from: http://www.ahpf.org.uk/AHP_Advisory_Fitness_for_Work_Report.htm [accessed 19th January 2014].

7. STURESSON, M., EDLUND, C., FJELLMAN-WIKLUND, A., FALKDAL, occupational therapists and physicians. Work, 2013; (45): pp117-128.

8. SCHONSTEIN, E. & KENNY, D. T. The value of functional and workplace assessments in achieving a timely return to work for workers with back pain. Work, 2001; (16): pp31-38.

9. GIBSON, L., & STRONG, J. Expert review of an approach to functional capacity evaluation. Work, 2002; (19): pp231-242.

10. STRONG, S. Functional Capacity Evaluations: the good the bad and the ugly. OT Now, 2002; 5-9.

11. OESCH, P. R., KOOL, J. K., BACHMANN, S. & DEVEREUX, J. The influence of a functional capacity evaluation on fitness for work certificates in patients with non-specific chronic low back pain. Work, 2006; (26): pp259-271.

12. JAMES, C. & MACKENZIE, L. The clinical utility of functional capacity evaluations: the opinion of health professionals working within occupational rehabilitation. Work, 2009; (33): pp231-239.

13. STRONG, S., BAPTISTE, S., CLARK, J., COLE, D. & COSTA, M. Use of Functional Capacity Evaluations in workplaces and the compensation system: a report on workers and report users perceptions. Work, 2004; (23): pp67-77.

14. INNES, E., & STRAKER, L. Reliability of work related assessments. Work, 1999; (13): pp107-124.

15. MCFADDEN, S., MACDONAL, A., FOGARTY, A., LE, S. & MERRITT, B.K. Vocational assessment: a review of the literature from an occupation-based perspective. Scandinavian Journal of Occupational Therapy, 2010; (17): pp43-48.

16. GOUTTEBARGE, V., WIND, H., KUIJER, P. P. F.M. & FRINGS-DRESEN, M. H. W. Reliability and validity of Functional Capacity Evaluation methods: a systematic review with reference to Blankenship system, Ergos work simulator, Ergo-Kit and Isernhagen work system. International Archive Occupational Environmental Health, 2004; (77): pp527-537.

17. VALPAR INTERNATIONAL CORPORATION, Joule: a comprehensive Industrial rehab system by Valpar training manual. Minneapolis: Valpar International Corporation; 2007.

18. VALPAR INTERNATIONAL CORPORATION, 1999. Inter-rater reliability of Joule: an FCE system by Valpar. [online]. USA: Valpar International Corporation, 1999-2011. Available from: http [accessed 17th February 2011].

19. CANADIAN SOCIETY FOR EXERCISE PHYSIOLOGY. The physical activities readiness questionnaire (PAR-Q). [online]. Canada: Canadian Society for Exercise Physiology; 2002. Available from: http://www.csep.ca/english/view.asp?x=698 [accessed 24th April 2011].

20. LEGGE, J. & BURGESS-LIMERICK, R. Reliability of the Job-fit System pre-employment functional assessment tool. Work, 2007; (28): pp299-312.

21. RENEMAN, M. F., BROUWER, S., MEINEMA, A., DIJKSTRA, P.U., GEERTZEN, J.H.B. & GROOTHOFF, J.W. Test-retest reliability of the Isernhagen Work Systems Functional Capacity Evaluation in healthy adults. Journal of Occupational Rehabilitation, 2004; 14(4): pp295-305.

22. BORG, G. Perceived exertion as indicator of somatic stress. Scandinavian Journal of Rehabilitative Medicine, 1970; 2(2): pp92-98.

23. DURAND, M., LOISEL, P., POITRAS, S., MERCIER, R., STOCK, S. R. & LEMAIRE, J. The inter-rater reliability of a functional capacity evaluation: the Physical Work Performance evaluation. Journal of Occupational Rehabilitation, 2004; 14 (2): pp119-129.

24. Data Protection Act, 1998. C. 29.

25. IBM CORPORATION, 2010. SPSS statistics 19 [online]. New York: IBM Corporation; 2010. Available from: http://www.spss.com/software/statistics/ [accessed 24th April 2011].

26. FIELD, A. Discovering Statistics Using SPSS 3rd Ed. London: Sage Publications Ltd; 2011.

27. KOTTNER, J., AUDIGE, L., BRORSON, S., DONNER, A., GAJEWSKI, B. J., HROBJARTSSON, A., ROBERTS, C., SHOUKRI, M. & STREINER, D. L. Guidelines for reporting reliability and agreement studies were proposed. Journal of Clinical Epidemiology, 2011; (64): pp96-106.

28. ELIASZIW, M., YOUNG, S. L., WOODBURY, M. G., & FRYDAY-FIELD, K. Statistical methodology for the concurrent assessment of Inter-rater and intra-rater reliability: using goniometric measurements as an example. Physical Therapy, 1994; (74): pp777-788

29. BRUTON, A., CONWAY, J.H., & HOLGATE, S.T. Reliability: What is it, and how is it measured? Physiotherapy, 2000; 86(2): pp.94-99.

30. GROSS, D. P., & BATTIE M. C. Reliability of safe maximum lifting determinations of a functional capacity evaluation. Physical Therapy, 2002; 82 (4): pp364-371.

31. JAMES, C., MACKENZIE, L. & CAPRA, M. Test-retest reliability of the manual handling component of the WorkHab Functional Capacity Evaluation in healthy adults. Disability Rehabilitation, 2011; 32(22): pp1863-1869.

32. GARDENER, L., & MCKENNA, K. Reliability of occupational therapists in determining safe, maximal lifting capacity, Australian Occupational Therapy Journal, 1999; (46): pp110-119.

33. KING, P. M., TUCKWELL, N. & BARRETT, T. E. A critical review of FCEs, Physical Therapy, 1998; 78(8): pp852-866.

34. CRONIN, S., CURRAN,J., IANTORNO, J., MURPHY, K., SHAW, L., BUTCHER, N. & KNOTT, M. Work capacity assessment and return to work: a scoping review. Work, 2013; (44): pp37-55.

35. TENGLAND, P. A qualitative approach to assessing work ability. Work, 2013; (44); pp393-404

36. TUCKWELL, N. L., STRAKER, L. & BARRETT, T. E. Test-retest reliability on nine tasks of the Physical Work Performance Evaluation. Work, 2002; (19): pp243-253.

37. ALLAN, J. L., JAMES, C. & SNODGRASS, S. J. The effect of load on biomechanics during overhead lift in the WorkHab Functional Capacity Evaluation. Work. 2012; (43): pp487-496.

38. STEMLER, S. E. A comparison of consensus consistency and measurement approaches to estimate inter-rater reliability. [online] Practical assessment, Research and Evaluation, 2004; 9(4): Available from: http://pareonline.net/getvn.asp?v=9&n=4 [accessed 1st September 2012].

39. BOWLING, A. Research Methods in Health Investigating Health and Health Services, 3rd Ed. New York: Open University Press; 2009.

40. GOUTTEBARGE, V., WIND, H., KUIJER, P. P., SLUITER, J. K. & FRINGS-DRESEN, M. H. Reliability and agreement of 5 Ergo-Kit Functional Capacity Evaluation lifting tests in subjects with low back pain. Archives of Physical Medical Rehabilitation, 2006; (87): pp1365-1370.

41. GOUTTEBARGE, V., WIND, H., KUIJER, P. P., SLUITER, J. K. & FRINGS-DRESEN, M. H. Intra- and inter-rater reliability of the Ergo-Kit Functional Capacity Evaluation Method in adults without musculoskeletal complaints. Archives of Physical Medical Rehabilitation, 2005; (86): pp2354-2360.

Table 1 : Intraclass Correlation Coefficient (ICC) and percentages of agreement for forceful lifting and carry tasks for all 3 raters

	ICC (95% CI)	Absolute agreement	Absolute disagreement
All forceful tasks		20.8% (15/72)	15.2% (11/72)
Waist to waist lift	.988 (0.917-0.998)	16.6% (2/12)	8.3% (1/12)
Waist to floor lift	0.988 (0.939-0.997)	8.3% (1/12)	33.3% (4/12)
Waist to above shoulder lift	0.973 (0.821-0.993)	8.3% (1/12)	16.6% (2/12)
Bilateral carry	.978 (0.930-0.993)	50% (6/12)	16.6% (2/12)
Unilateral dominant carry	.950 (0.738-0.987)	8.3% (1/12)	8.3% (1/12)
Unilateral non-dominant carry	.933 (0.835-0.978)	33.3% (4/12)	8.3% (1/12)

Key: ICC – intraclass correlation; CI – confidence interval
Forceful Tasks: 6 per 12 forceful task for each subject (n=12)= total 72

Table 2: Intraclass Correlation Coefficient (ICC), confidence intervals, significance and percentages of agreement for forceful lifting and carry tasks for paired raters

Forceful Task	Paired Raters 1 and 2			Paired Raters 2 and 3			Paired Raters 1 and 3		
	% agreement	Paired sample ICC (95% CI)	Sig (2 tailed) p=	% agreement	Paired sample ICC (95% CI)	Sig (2 tailed) p=	% agreement	Paired sample ICC (95% CI)	Sig (2 tailed) p=
Waist to waist lift	25% (3/12)	.995 (-1.341 - -0.492)	.001	16.6% (2/12)	.993 (0.580 - 1.587)	.001	83.3% (10/12)	.998	.166
Waist to floor lift	33.3% (4/12)	.995 (-0.345 - -3.527)	.005	16.6% (2/12)	.995 (-0.161 - 0.828)	.166	33.3% (4/12)	.995 (-0.161 - 0.828)	.166
Waist to above shoulder lift	16.6% (2/12)	.990 (-1.383 - -0.617)	.000	16.6% (2/12)	.990 (0.617 - 1.383)	,000	66.7% (8/12)	.990 (-0.383 - 0.383)	1.000
Bilateral carry	58.3% (7/12)	.984 (-1.857 - -0.143)	.026	50% (6/12)	.985 (0.317 - 2.016)	.012	75% (9/12)	.990 (-0.488 - 0.821)	.586
Unilateral dominant carry	16.6% (2/12)	.988 (-1.244 - -0.589)	.000	33.3% (4/12)	.981 (0.354 - 0.980)	.001	58.3% (7/12)	.980 (-0.645 - 0.145)	.191
Unilateral non-dominant carry	50% (6/12)	.974 (-0.832 - -0.168)	.007	50% (6/12)	.909 (-0.645 - 0.145)	.586	58.3% (7/12)	.949 (-0.828 - 0.161)	.166
All forceful tasks	33.3% (24/72)			30.5% (22/72)			62.5% (45/72)		

Key: ICC – intraclass correlation; CI – confidence interval; Forceful Tasks: 6 forceful tasks for each subject (n=12)= total 72

Table 3: percentage of rater agreement for reasons for terminating forceful, repetitive and positional tasks

Task	Rating test fully completed			Rating unsafe body mechanics			Rating subject requests stop			
Task	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	% agreement
Forceful tasks										
Waist to waist lift	0	0	0	12	12	12	0	0	0	100 (12/12 subjects)
Waist to floor lift	0	0	0	12	12	12	0	0	0	100 (12/12 subjects)
Waist to above shoulder	0	0	0	12	12	12	0	0	0	100 (12/12 subjects)
Bilateral carry	0	0	0	12	12	12	0	0	0	100 (12/12 subjects)
Unilateral dominant hand carry	0	0	0	12	12	12	0	0	0	100 (12/12 subjects)
Unilateral non-dominant hand carry	0	0	0	12	12	12	0	0	0	100 (12/12 subjects)
Repetitive tasks										
walking	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Balance	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Ladder climb	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Stair climb	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Repetitive squat	10	10	11	1	1	0	1	1	1	91.7 (11/12 subjects)
Repetitive foot -right	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Repetitive foot -left	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Crawl	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Positional Tasks										
Kneel	10	10	11	0	0	0	2	2	1	91.7 (11/12 subjects)
Crouch	8	8	9	1	1	0	3	3	3	91.7 (11/12 subjects)
Midlevel reach	11	11	11	0	0	0	1	1	1	100 (12/12 subjects)
Elevated reach	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Sit	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Stand	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)

6 forceful, 8 repetitive and 6 positional tasks per subject (n=12)

Table 4: percentage of rater agreement for scoring maximum safe capacity in repetitive and positional tasks

	Rating No Limitations			Rating Occasional			Rating Rare			
Task	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	Rater 1	Rater 2	Rater 3	% agreement
Repetitive tasks										
walking	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Balance	11	11	11	1	1	1	0	0	0	100 (12/12 subjects)
Ladder climb	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Stair climb	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Repetitive squat	10	10	11	2	2	1	0	0	0	91.7 (11/12 subjects)
Repetitive foot -right	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Repetitive foot -left	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Crawl	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Positional Tasks										
Kneel	10	10	11	2	2	1	0	1	0	91.7 (11/12 subjects)
Crouch	8	8	9	3	3	2	1	1	1	91.7 (11/12 subjects)
Midlevel reach	11	11	11	1	1	1	0	0	0	100 (12/12 subjects)
Elevated reach	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Sit	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)
Stand	12	12	12	0	0	0	0	0	0	100 (12/12 subjects)

8 repetitive and 6 positional tasks per subject (n=12)